

Many researchers would be concerned if they learned that some core conclusion of their statistical analysis—such as the sign or statistical significance of some key effect—could be overturned by removing a small fraction of their data. Such non-robustness would be particularly concerning if the data were not actually drawn randomly from precisely the population of interest, or if the model may have been misspecified—circumstances that often occur in the social sciences. For example, a study of microcredit in Mexico on 16,561 households measured a negative but statistically insignificant effect of microcredit [Angelucci et al., 2015]. However, in recent work, my co-authors and I show that one can make the estimated effect of microcredit positive and statistically significant by removing just 15 households from the analysis, reversing the paper’s qualitative conclusions by removing only 0.1% of the data. Since there are a combinatorially large number of ways to leave 15 datapoints out of 16,561 (over 10^{51}), finding such influential subsets by brute force would be impossible. I circumvent this difficulty by forming a *linear approximation* to the dependence of the estimator on the dataset. My technique works for a wide class of commonly used estimators, providing a fast, automatic tool for identifying small but influential subsets with finite-sample guarantees in many applied problems.

In fact, my research shows that many desirable but computationally demanding data analysis tasks are similarly amenable to automatic approximation, providing fast, easy-to-use computational tools for assessing whether small changes to modeling assumptions or datasets can meaningfully alter an analyst’s substantive conclusions. As I detail below, my work encompasses Bayesian approaches (such as measuring robustness to prior specification), frequentist approaches (such as fast approximations to leave-one-out cross validation), and combinations of the two (such as measuring the robustness of Markov chain Monte Carlo procedures to data resampling). For the remainder of this statement, I will describe some of my specific past and future projects, emphasizing the ways in which I update classical results for a modern computational environment using intuitive, relevant theory, and with motivation drawn from practical applications.

Approximate cross validation. The error or variability of machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data; cross-validation (CV) can be thought of as a particularly popular example of this technique. In Giordano et al. [2019b], I use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster by an order of magnitude than repeated re-fitting. I provide explicit finite-sample error bounds for the approximation in terms of a small number of simple, verifiable assumptions. My results apply whether the weights and data are stochastic or deterministic, and so can be used as a tool for proving the accuracy of the approximation on a wide variety of problems. As a corollary, I state mild regularity conditions under which the approximation consistently estimates true leave- k -out cross-validation for any fixed k . I demonstrate the accuracy of the approximation on a range of simulated and real datasets, including an unsupervised clustering problem from genomics.

Prior sensitivity for discrete Bayesian nonparametrics. A central question in many probabilistic clustering problems is how many distinct clusters are present in a particular dataset and which observations cluster together. Discrete Bayesian nonparametric (BNP) mixture models address this question by placing a generative process on cluster assignment, making the number and composition of distinct clusters amenable to Bayesian inference. However, like all Bayesian approaches, BNP requires the specification of a prior, and this prior may favor different numbers and types of posterior clusters.

In Giordano* et al. [2021b], I derive and analyze prior sensitivity measures for variational Bayes (VB) approximations in general, with a practical focus on discrete BNP models. Unlike much previous work on local Bayesian sensitivity for BNP, I pay special attention to the ability of the sensitivity measures to *extrapolate* to different priors, rather than treating the sensitivity as a measure of robustness *per se*. I state conditions under which VB approximations are Fréchet differentiable functions of prior densities in a particular vector space, while also proving that VB approximations are in fact *non-differentiable* in another wide class of vector space embeddings popular in the classical Bayesian robustness literature. My co-authors and I apply the sensitivity measures to a number of real-world problems, including an unsupervised clustering problem from genomics using fastSTRUCTURE. We demonstrate that the approximation is accurate, orders of magnitude faster than re-fitting, and capable of detecting meaningful prior sensitivity in quantities of practical interest.

Uncertainty propagation in mean-field variational Bayes. Mean-field Variational Bayes (MFVB) is an approximate Bayesian posterior inference technique that is increasingly popular due to its fast runtimes

on large-scale scientific data sets. For example, in Regier et al. [2019], my co-authors and I use MFVB to construct an approximate posterior for the identity of every astronomical object in 55TB of image data from the Sloan Digital Sky Survey. However, even when MFVB provides accurate posterior means for certain parameters, it often mis-estimates variances and covariances due to its inability to propagate Bayesian uncertainty between statistical parameters.

In Giordano et al. [2015, 2018a], I derive a simple formula for the effect of infinitesimal perturbations on MFVB posterior means, thus providing improved covariance estimates and greatly expanding the practical usefulness of MFVB posterior approximations. My approach builds on a result from the classical Bayesian robustness literature relating posterior covariances to the derivatives of posterior expectations, and includes the classical Laplace approximation as a special case. In experiments on simulated and real-life datasets, including models from ecology, the social sciences, and on a massive internet advertising dataset, I demonstrate that my method is simple, general, and fast, providing accurate posterior uncertainty estimates and robustness measures with runtimes that can be an order of magnitude faster than MCMC.

The empirical influence function. Much of my work (particularly Giordano et al. [2019b], Broderick et al. [2020], Giordano and Broderick [2021]) has strong connections to the classical theory of von Mises expansions and the closely related concept of the influence function, which measures the effect of individual datapoints on an estimator [Mises, 1947, Hampel, 1986, Serfling, 2009]. But my focus on the influence function evaluated at the observed data—i.e., the “empirical influence function” (EIF)—stands in contrast with much of the classical literature, which studies the asymptotic behavior of estimators via their (unobserved) limiting influence function. In our present age of automatic differentiation, large datasets, and complex models, I believe that the EIF will continue to provide practical benefits and is relatively under-studied.

In Giordano et al. [2019a], I show that higher-order EIFs can be easily and automatically evaluated and analyzed for M-estimators at a computational cost comparable to the first-order EIF—that of forming and factorizing a Hessian matrix of second-order derivatives. Thus, the EIF “amortizes” the cost of evaluating an M-estimator large number of alternative datasets: by paying a large fixed price up front (approximately computing and factorizing a Hessian matrix), one can cheaply approximate M-estimators at a very large number of alternative datasets. Natural applications of the idea include approximating the bootstrap-after-bootstrap, evaluating the sampling properties of cross-validation, and computing higher-order jackknife bias correction.

Sensitivity analysis in difficult situations. It is not always as easy to apply sensitivity analysis in practice as it is in theory. I have found that a few key problems tend to recur, and I will discuss them in turn, as well as potential solutions which draw connections to the optimization literature.

First, sensitivity analysis should, ideally, deal gracefully with incomplete optimization. For example, a collaborator from biostatistics and I have found that the popular R package `DESeq2` can fail in practice to fully optimize the log likelihood, and so fail to satisfy the assumptions that make sensitivity analysis possible. Instead of forcing users to optimize further, I propose that second-order empirical influence functions could simulate the effect of simultaneously taking a Newton step and perturbing the data, permitting sensitivity analysis on incompletely optimized objectives with little computation beyond that required for well-optimized objectives.

Second, the key computational bottleneck in sensitivity analysis in high-dimensional problems is typically the solution of linear systems involving the inverse Hessian of the objective function. Off-the-shelf iterative algorithms like the conjugate gradient algorithm suffice in many cases, but there is reason to believe that the present active research into stochastic second-order methods could significantly speed up sensitivity analysis in large problems.

Finally, practitioners are often interested in non-smooth objectives. For these, one promising idea is to use local approximations to speed up computationally intensive but smooth components in non-smooth problems. For example, Wilson et al. [2020] speeds up cross-validation of linear regression with a non-smooth lasso penalty by forming a fast approximation to the effect on the optimal squared error of leaving out a single datapoint, and retaining non-smoothness in the lasso penalty. We take a similar approach in Giordano* et al. [2021b], retaining easy-to-compute non-linearities in posterior summary statistics.

References

- M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.
- T. Broderick, R. Giordano*, and R. Meager*. An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*, 2020. *Following conventions in econometrics, the authors are listed alphabetically. Giordano and Meager are equal contribution primary authors.
- R. Giordano and T. Broderick. The Bayesian infinitesimal jackknife for variance. *In preparation.*, 2021.
- R. Giordano, T. Broderick, and M. I. Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449, 2015.
- R. Giordano, T. Broderick, and M. Jordan. Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029, 2018.
- R. Giordano, M. I. Jordan, and T. Broderick. A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019a.
- R. Giordano, W. Stephenson, R. Liu, M. I. Jordan, and T. Broderick. A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147, 2019b.
- R. Giordano*, R. Liu*, M. I. Jordan, and T. Broderick. Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. <https://arxiv.org/abs/2107.03584>, 2021. *Giordano and Liu are equal contribution primary authors.
- F. Hampel. *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience, 1986.
- R. Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.
- J. Regier, K. Fischer, K. Pamnany, A. Noack, J. Revels, M. Lam, S. Howard, R. Giordano, D. Schlegel, and J. McAuliffe. Cataloging the visible universe through Bayesian inference in Julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104, 2019.
- R. Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.
- A. Wilson, M. Kasy, and L. Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR, 2020.